Study of an automated procedure for a Dutch sentence test for the measurement of the speech reception threshold in noise

Hayo Terband[a] and Rob Drullman
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

(Received 17 May 2007; revised 20 August 2008; accepted 27 August 2008)

A procedure was developed for the automated measurement of the speech reception threshold in stationary noise (SRTn), which can be administered by the subjects themselves using a computer. The procedure was based on the SRTn test for Dutch developed by Plomp and Mimpen [(1979). "Improving the reliability of testing the speech reception threshold for sentences," Audiology, **18**, 43–52]. Because in the automated procedure the responses were entered on a keyboard, the question of how to deal with typing and spelling errors played a key role. At first the possibility of scoring on keywords only was examined. An experiment was conducted in which the adaptive procedure was varied. Results showed that the combination of scoring each keyword separately and a fixed scheme of the adaptation of the signal-to-noise ratio throughout the procedure yields the highest test-retest reliability. Subsequently, the collection and verification of responses using a keyboard were examined. Two different algorithms were developed and evaluated against the traditional task of verbal repetition and response verification by an experimenter. The results indicated a preference for verification by dynamic alignment over a spelling checker approach. In conclusion, the results show that it is possible to automate the test procedure while maintaining sufficient reliability.
 [DOI: 10.1121/1.2990706]

PACS number(s): 43.71.Gv, 43.71.Ky, 43.71.Es [DOS] Pages: 3225–3234

I. INTRODUCTION

Listeners with hearing loss generally consider the understanding of speech in noisy environments to be the most troublesome aspect of their impairment (Wagener *et al.*, 2003; Smits *et al.*, 2004). The ability to understand speech in noise can be quantified as the signal-to-noise ratio (SNR) necessary to reach a 50% intelligibility score, which is denoted as the speech reception threshold in noise (SRTn) [e.g., see Plomp and Mimpen (1979), Letowski *et al.* (1992), and Noordhoek *et al.* (1999, 2000)]. To measure this ability, SRTn tests have been developed for a wide variety of languages. Most of these tests consist of a repetition task of a simple straightforward sentence material that is presented against a background of (stationary speech-shaped) noise. Although this testing method has proved to be generally very accurate and reliable [e.g., see Smoorenburg (1992), Hagerman and Kinnefors (1995), Kollmeier and Wesselkamp (1997), and Versfeld *et al.* (2000)] it has the disadvantage that an experimenter is needed to administer the test and to decide whether a sentence is repeated correctly or not. The elimination of this disadvantage requires an automatic test procedure, and several automatic tests have been developed. For example, the telephone screening test for Dutch, with simple three-digit stimuli, proved quite successful in its domain (Smits *et al.*, 2004). In the present paper, we present the development of a fully automated *sentence* test that subjects can take independently on a computer.

For Dutch, the test developed by Plomp and Mimpen (1979) is considered to be the standard (Smoorenburg, 1992; Versfeld *et al.*, 2000; Smits *et al.*, 2004), and this test is widely used for both clinical and research purposes [e.g., see Plomp (1986), Rooij and Plomp (1991), and Noordhoek *et al.* (1999, 2000)]. The Plomp and Mimpen test consists of a list of 13 sentences of eight to nine syllables that are presented in a background of speech-shaped noise. The noise level is fixed at a comfortable level. The level of the sentences is varied according to an adaptive procedure. A sentence is considered to be repeated correctly only if the whole sentence is reproduced verbatim. The first sentence of a list is presented at a SNR of −8 dB, which is normally below the reception threshold. This sentence is repeated, each time at a 4 dB higher SNR until the subject can reproduce it correctly. The remaining 12 sentences are each presented once in a simple up-down procedure with a 2 dB step size. The speech reception threshold is calculated by averaging the presentation levels of sentences 5–13 and the level at which a 14th sentence would have been presented (Plomp and Mimpen, 1979).

The goal of our study was to design an automatic procedure based on the existing speech material, with a test-retest reliability similar to the standard SRTn-test by Plomp and Mimpen (1979). The first question that arose when considering the automation of a SRTn-test was whether, and if so how, the test method needed to be adapted. It might not be possible to maintain the same test procedure, and automation might introduce factors that affect the test validity. On the other hand, automation might enable methodological adjustments that improve the test validity. In an automatic proce-

[a]Present address: ENT/Medical Psychology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.

dure, the responses are entered by the subject himself/herself using the keyboard. In this paper we present two experiments regarding the development of such a procedure. The first experiment was set up in order to evaluate the effect of scoring just the keywords of each sentence instead of scoring each sentence as a whole. In the second experiment, two approaches of collecting and verifying keyboard responses were evaluated.

The nature of scoring (scoring whole sentences as a block, scoring all words separately, scoring only keywords, etc.) is known to have a range of effects on SRTn measurements. A comparison of speech reception tests indicates that scoring all words separately, instead of scoring the sentence as a whole, results in increased measurement efficiency, often enabling quicker measurements (Brand and Kollmeier, 2002). This is in accordance with the view of Hagerman and Kinnefors (1995) that fewer scoring units (whole sentences versus words) probably cause a higher standard deviation of individual SRTn values in repeated measurements. On the other hand, the sentence material of the Plomp and Mimpen test was not originally intended to be used with keyword or separate word scores. The test is based on balanced lists for optimal whole sentence scoring. Versfeld *et al.* (2000) reported a decrease in the slope of the discrimination function of 1.3%/dB when using the number of correctly repeated words instead of whole sentence scoring. In view of our objective to create an automatic procedure, we anticipated that a large set of keywords would provide the best basis. In that way, as much sentence information as possible is used, combined with a sufficient number of scoring units.

Apart from the use of keyword scoring, the adaptive procedure for reaching the final SRTn was also investigated in the first experiment. That is, we evaluated different strategies for the scoring of keywords and for the step size of the SNR adaption (viz., increasing or decreasing the SNR of the next sentence). The Plomp and Mimpen test uses an open set of sentences. As the number of words and keywords per sentence varies, scoring (key) words separately becomes more complicated with an open set of sentences. However, relatively few of the open-set tests score whole sentences as a block. In fact only the tests by Plomp and Mimpen (1979), Gelfland *et al.* (1988), and Nilsson *et al.* (1994) do. Other open-set SRTn-tests, such as the speech perception in noise (SPIN)-tests (Kalikow *et al.*, 1977; Bilger *et al.*, 1984) and the Göttingen sentence test (Kollmeier and Wesselkamp, 1997), calculate scores in percentage correct or only score a sentence correct if all *keywords* are repeated correctly [e.g., the open-set test by Macleod and Summerfield (1990)]. In our experiment we compared scoring per keyword (relative scoring) with scoring the entire set of keywords (block scoring).

On the basis of the scoring result, the step size of the SNR adaptation is determined when proceeding through the list of sentences. Most tests that score sentences as a block use fixed adaptation steps of ±2 dB [e.g., see Plomp and Mimpen (1979), Macleod and Summerfield (1990), and Nilsson *et al.* (1994)]. When (key)words are scored separately, more complicated stepping schemes can be developed. For example, the method used by Hagerman and Kinnefors (1995) consists of a fixed stepping scheme ranging from −2 to +3 dB in 1 dB steps, depending on the number of words repeated correctly. In our experiment we used a variant of this fixed step size scheme and also evaluated a more complex paradigm as proposed by Brand and Kollmeier (2002), where the step size effectively decreases as the sentence number in a list increases (gradual convergence toward the SRTn).

The second experiment focused on collecting responses using a computer keyboard (instead of verbal responses) and the problem of dealing with spelling and typing errors. A SRTn-test is supposed to measure speech reception and not the subject's spelling or typing abilities. How can one know whether something is a spelling or typing error and not a result of misperception? The obvious solution was to find a balance in which the errors that are likely to be spelling or typing errors are discarded, while errors that are probably due to misperception are penalized. By using a keyword approach from the first experiment, the influence of typing errors on the test results was reduced and the performance of the recognition algorithm was maximized. Two algorithms were compared: one simple spelling checker and a more complex dynamic alignment paradigm (Wagner and Fisher, 1974).

The effectiveness or success of the final automated procedure and of the intermediate results was determined by comparison with the standard Plomp and Mimpen test. This was done in terms of the estimated SRTn value, the test-retest reliability, and the shape/slope of the discrimination function. Before describing the two experiments, the selection of keywords and the adjustments to the adaptive procedures are discussed in the following section.

II. ADAPTIVE PROCEDURES

A. Selection of keywords

A first step in automating the SRTn procedure was to make use of keyword scoring. When deciding what set of keywords should be used, two approaches were possible. The first approach (*bottom-up*) considered the words that should be included. This set started with all content words, the "true" keywords (e.g., nouns and adjectives), and was enlarged by all function words that clearly had an important semantic function within the sentence context. The second approach (*top-down*) started with the whole sentence and checked which words could be left out. These were usually articles and (personal) pronouns with a less important semantic function. In practice the application of the two approaches led to virtually the same result, corresponding with the extraction of "newspaper headlines" from the sentences. The number of keywords ranges from 3 to 5, with an average of 3.7 keywords per sentence.[1]

B. Scoring and step size

With a limited number of keywords per sentence, the decision of a correct score can be made in two ways: either all keywords must be correct (block scoring) or scores are given for each keyword that is reproduced correctly (relative scoring). Referring to Hagerman and Kinnefors (1995), we

TABLE I. Stepping scheme for keyword scoring for sentences containing three to five keywords (e.g., a score of 2/5 means two out of five keywords correct).

Score	0/5	1/5	1/4	1/3	2/5	2/4	3/5	2/3	3/4	4/5	5/5
Step (dB)	+2	+2	+1.5	+1.5	+1	0	−1	−1.5	−1.5	−2	−2

considered a fixed stepping scheme based on the number of keywords that are repeated correctly. We developed a symmetrical stepping scheme for measuring the 50% threshold (Table I). A complicating factor was the fact that in the Plomp and Mimpen test the number of keywords per sentence is not fixed but varies from 3 to 5. With a 1.5 dB step size for the 1/4, 1/3, 2/3, and 3/4 scores, the five step scheme shows a balanced distribution of step sizes.[2] Note that in the block scoring condition this scheme results in a fixed step size of ±2 dB since in this condition the score is either 0 or 1.

Brand and Kollmeier (2002) developed a generalization of the Hagerman and Kinnefors method in which the change in the presentation level of the subsequent sentence is calculated on the basis of the following variables: *percentage of correct (key)words of the previous sentence (prev), number of reversals of the presentation level (i)*, and constant *target discrimination value (tar)*, and *slope of the discrimination function (slope)*. In formula,

$$\Delta L = -\frac{f(i)(\text{prev-tar})}{\text{slope}}. \quad (1)$$

The function $f(i)$ returns a smaller number as i increases; as a result the step size decreases as the list proceeds. This procedure is used either with an open or a closed set of sentences [the Göttingen and the Oldenburger sentences, see Kollmeier and Wesselkamp (1997) and Wagener *et al.* (1999a, 1999b, 1999c)] and could be implemented for the Plomp and Mimpen test simply by filling in the constants. Tar was set to 0.5 because the Plomp and Mimpen test measures the 50% speech reception threshold. The slope parameter was set to 0.15 dB^{-1}, in accordance with the slope of the discrimination function found by Plomp and Mimpen (1979).

In summary, two different scoring strategies were investigated: *block scoring* (all keywords or nothing) and *relative scoring* (number of correct keywords re all keywords). Furthermore, two different step size strategies were investigated: *fixed stepping scheme* (Hagerman and Kinnefors style) and *dependent step size* (Brand and Kollmeier style).

III. EXPERIMENT 1: COMPARISON OF ADAPTIVE PROCEDURES

A. Methodology and materials

1. Stimuli and design

The original test material as developed by Plomp and Mimpen (1979) was used, consisting of ten lists of 13 sentences spoken by a female speaker. All sentences were available in digital form with a sampling rate of 44,100 Hz and a 16 bit resolution. The keyword strategy was evaluated in four conditions (two scoring conditions × two step size conditions) against the standard procedure, which makes a total of five test conditions.

2. Subjects

A total of 15 subjects with normal hearing participated in the listening experiment. Normal hearing was defined as having a pure-tone threshold not exceeding 15 dBHL at any frequency from 250 to 2000 Hz and not exceeding 25 dBHL in the range of 2000–8000 Hz for both ears. The better ear was used for testing. The subjects might have experience with speech-in-noise testing, but none of the subjects were familiar with the Plomp and Mimpen sentence material.

3. Procedure

The ten lists were presented in a fixed order. Each of five conditions was tested with two lists. The conditions were tested in alternate orders, according to a 5×5 latin-square design. Having 15 subjects, each sequence was presented to 3 subjects. Three lists of sentences from the female speaker of Versfeld *et al.* (2000) were presented as practice materials.

The signal was delivered monaurally by headphones (Sony MDR-7509) through a RME Hammerfall DSP Multiface sound module. The noise level was fixed at a comfortable level of 70 dBA. The mixing of the speech and noise signal was done by a computer program according to the adaptive procedure under test. The subjects received written instructions. The subjects' task was to repeat the sentences verbally. As in the original test, each first sentence of a list was presented at a SNR of −8 dB. This sentence was repeated, each time at a 4 dB higher level, until the subject could reproduce it correctly (in the case of the keyword conditions, i.e., all keywords correct). The remaining 12 sentences were presented once in an up-down procedure according to the strategy under test. The tests were administered by an experimenter, who evaluated the responses.

4. Data analysis

The test-retest reliability was defined in accordance with Plomp and Mimpen (1979), viz., as the root mean square of the differences between the two SRTn values for each subject in each condition, divided by $\sqrt{2}$ [see also Nilsson *et al.* (1994), Hagerman and Kinnefors (1995), Versfeld *et al.* (2000), Brand and Kollmeier (2002)]. A permutation test was performed to check whether the differences among conditions were significant. For each condition 20 permutations of 10 out of 15 were drawn from the original data. (The number 10 was chosen because this is the number of subjects used by Plomp and Mimpen (1979) in their original evaluation.) For each of these samples, the test-retest reliability was estimated. Subsequently, an analysis of variance was performed on the 5×20 resulting values.

The discrimination function results from plotting the proportion of correct responses over all subjects for all presentation levels, thus indicating the intelligibility score as a function of presentation level. Because the number of sentences per individual subject was too small for a slope estimate, all data were pooled across subjects [see Plomp and

TABLE II. Mean SRTn (standard deviation in parentheses), test-retest reliability, and slope of the discrimination function of SRTn measurements for the five test conditions of experiment 1. Mean SRTn refers to the value across presentation levels; ML SRTn refers to the value obtained with the ML estimation.

Condition	Mean SRTn (SD) (dB)	Test-retest reliability (dB)	Slope (%/dB)	ML SRTn (dB)
0: standard test	−5.63 (0.86)	0.75	21	−5.65
1: relative/fixed	−6.99 (0.87)	0.69	15	−6.76
2: block/fixed	−5.83 (1.16)	1.17	16	−5.76
3: relative/dependent	−7.12 (0.86)	0.86	16	−7.13
4: block/dependent	−5.61 (1.45)	1.54	14	−5.39

Mimpen (1979)]. Note that this required that all subjects behaved in the same manner. In the block scoring conditions, the proportion simply equaled the relative number of sentences reproduced correctly. In the relative scoring conditions, the proportion of correct responses was obtained by averaging the relative number of keywords that was reproduced correctly for each sentence presented at each presentation level. The function was fitted by the maximum-likelihood (ML) criterion as developed by Versfeld *et al.* (2000), in which the SRTn and the slope are estimated in an iterative procedure. Besides a slope value, the ML estimation of the discrimination function also produces an estimate of the SRTn.

B. Results

Mean SRTn values in the five conditions are given in Table II. The mean values are based on the average of 30 individual SRTns per condition, as each subject was tested with two lists in each condition. The results show that the block scoring conditions (2 and 4) produce virtually the same SRTn as the standard test. The relative scoring conditions (1 and 3), on the other hand, seem to lead to lower SRTn values in comparison with the standard test. A repeated-measure analysis of variance indicates a significant difference among conditions ($p<0.01$). A subsequent Tukey-HSD *post hoc* test shows that both relative scoring conditions differ significantly ($p<0.01$) from the standard test, which indicates an effect of the scoring method on the SRTn value. An analysis of variance with *scoring* and *step size* as fixed factors (discarding data of the standard test) shows a significant effect of scoring ($p<0.01$), no effect of step size, and no interaction between factors.

To see if a learning effect occurred, the first and second measurements were compared for each subject in each condition. The results indicated no significant difference averaged over all conditions. To check whether any of the effects were individually significant, separate paired *t*-tests were performed per condition. No learning effect was observed.

In terms of test-retest reliability, the results in the relative scoring conditions are comparable to those of the standard test (values of 0.69 and 0.86 dB against 0.75 dB, Table II). With block scoring, the figures are considerably higher (1.17 and 1.54 dB). A repeated-measure analysis of variance

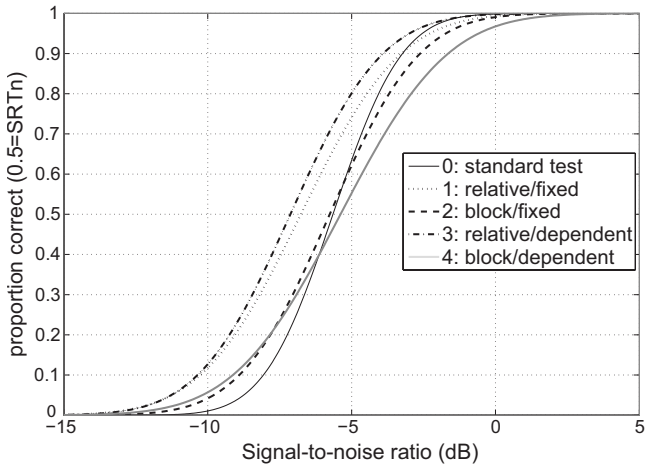

FIG. 1. Discrimination functions for the measurement of the SRTn in the five test conditions in experiment 1. See text and Table II for more details on slope values.

on the permutation resamples indicated a significant difference among conditions ($p<0.01$). A subsequent Tukey-HSD *post hoc* test, however, showed that the relative scoring conditions (1 and 3) did not differ significantly from the standard test.

To test whether pooling the data across subjects was justified for the estimation of the discrimination functions, an analysis of variance was conducted on the standard deviation within lists (i.e., the standard deviation of the mean score over the last ten sentences per list) with subject and condition as fixed factors. The results showed a significant effect of condition but no effect of subject and no interaction between subject and condition, thus indicating that all subjects behaved in the same manner. To see if a learning effect occurred, the first and second measurements were compared for each condition in a series of paired *t*-tests. No learning effect was observed. Therefore, all data were pooled across subjects for the estimation of the discrimination functions. Discrimination functions are shown in Fig. 1. Slope values and accompanying estimated SRTns (ML SRTn) are presented in Table II. Slope values range from 14%/dB (condition 4: relative/dependent) to 21%/dB (condition 0: standard test).

C. Discussion

1. SRTn

Reported common values (average SRTns for subjects with normal hearing) for the standard test (all words correct) range between −4.5 and −5.8 dB, depending on the sentence material [see Smits *et al.* (2004) for an overview]. The mean monaural SRTns found by Plomp and Mimpen (1979) are −5.6 dB (left ear only) and −6.2 dB (right ear only).

The results show that blocked keyword scoring produces the same SRTn as the standard test, whereas relative keyword scoring leads to a lower SRTn. This difference has an easy explanation: in block scoring any error directly leads to a zero score, where relative scoring often leads to a positive score. The effects on the progression of the adaptive procedure can be large. For example, three out of four keywords reproduced correctly yields a zero score in block scoring,

leading to an *increase* in the SNR for the next sentence, against a score of 0.75 in relative scoring, which is above the 50% threshold and leads to a *decrease* in SNR for the subsequent sentence. Hence, a block scoring threshold is different from a relative scoring threshold. For the sake of comparing the SRTn values with the standard test, however, SRTns acquired with the relative scoring procedures could be transformed by adding 1.4 and 1.5 dB for fixed and dependent step sizes, respectively.

As mentioned in Sec. III A 4, the ML estimation of the discrimination function, besides a slope value, also produces an estimation of the SRTn. The ML SRTn agrees well with the mean of the SRTns that were estimated by averaging across presentation levels, with a maximum difference of 0.23 dB (Table II).

2. Test-retest reliability

In terms of the standard deviation of individual SRTn values, the relative scoring conditions show results similar to the standard test. Among the keyword conditions, the highest test-retest reliability was found for the relative scoring condition in combination with a fixed stepping scheme (0.69 dB). Note that the figure we found for the standard test (0.75 dB) is lower than the standard deviation of 0.9 dB found by Plomp and Mimpen (1979). A closer look at the differences among the test conditions and the influences of the scoring factor indicates that relative scoring produces a higher test-retest reliability than block scoring. This confirms the prediction based on Hagerman and Kinnefors (1995) that scoring all keywords separately has a positive effect on the test reliability.

Regarding step size, however, the results are not in accordance with the original expectations. The data show a trend that the dependent step size conditions have a lower test-retest reliability than their fixed step size equivalents (Table II). The results, however, are in accordance with experience gained in practice. During the testing sessions, the adaptation of the presentation level in the dependent step size conditions repeatedly seemed to become too small, which resulted in an improbably high or low SRTn value for the given subject (based on the preceding and following measurements). The point is that although the dependent step size adaptive method reduces the standard deviation within measurements (i.e., the standard deviation of the presentation levels in single SRTn-measurements), it seems to produce less consistent SRTn estimates than the fixed step size method.

3. Discrimination function

The discrimination functions (Fig. 1) show that the standard test has the steepest slope of all test conditions. The slope value found for the standard test is 21%/dB, which is high, not only compared to the slope of 15%/dB found by Plomp and Mimpen for the exact same test,[3] but also compared to the slope values found in the literature for other tests [e.g., see Bilger *et al.* (1984), Gelfland *et al.* (1988), Nilsson *et al.* (1994), Kollmeier and Wesselkamp (1997), and Wagener *et al.* (2003)]. Among the other test conditions, the results show no large differences, slope values ranging from 14%/dB to 16%/dB.

For all experimental keyword conditions, the slope of the discrimination function is shallower than the slope that we found for the standard test. This may be due to the normalization of the sentence material. The rms levels of the sentences have been adapted so that each sentence became equally intelligible. Verifying only the keywords could interfere with the normalization because the rms of the keywords could differ from the rms of the whole sentence, and the size of the difference could vary among sentences. Consequently, if verified on keywords, sentences are no longer necessarily equally intelligible, which could have a negative effect on the test's discrimination function. The slope of the discrimination function of the tested methods might be improved by renormalizing the sentence material, which can be done by calculating the difference between the rms of the whole sentence and the rms of only the keywords for each sentence and by correcting the normalized sentence levels for the differences.

D. Summarizing remarks

The combination of relative scoring and a fixed stepping scheme produced the results that are the most similar to the standard test (condition 1 in Table II). The test-retest reliability is equivalent, but the slope of the discrimination function is shallower than the slope that we found for the standard test. However, our slope is comparable to other tests (Gelfland *et al.*, 1988; Wagener, 1999a, 1999b, 1999c; Versfeld *et al.*, 2000; Wagener *et al.*, 2003). Therefore, we used the combination of relative scoring with a fixed stepping scheme for further development of the automatic procedure.

IV. THE AUTOMATIC VERIFICATION OF RESPONSES

A. Verification procedures

The next step in automation is the verification of responses. We have explored two approaches to deal with this issue. The first approach is to use a spelling checker that consists of two steps: First, the spelling and typing errors are corrected; subsequently the responses are matched to the keywords in question. The outcomes of this approach depend on the dictionary and rules of the spelling checker. The second approach combines the aspects of correction and matching in one step of determining string similarity by dynamic alignment. Dynamic alignment, in its classical approach, is based on the minimal edit distance between the input string and a reference string (Wagner and Fisher, 1974; Yang *et al.*, 2003). For example, the deviation of the word *task* in reference to the word *token* can be expressed as a substitution (*a* for *o*), an insertion (*s*), and two deletions (*e* and *n*). In calculating the minimal edit distance, different deviations can be treated differently, so the effect of insertions, deletions, and substitutions can be individually regulated, and the likelihood of different substitutions can also be taken into account.

A more detailed description of the two methods is given in Sec. IV C. The first step in the development of verification

algorithms was to establish what kinds of errors occur when subjects have to enter their responses by keyboard. In order to do so, a pilot experiment was set up, which is presented below.

B. Pilot experiment: Error analysis

1. Methodology and materials

A total of seven subjects with no reported hearing problems and with varying computer experience and typing skills were asked to take an automatic test. Ages varied from 20 to 60 years. Each subject took the test repeatedly (varying from two to ten times), each time using different sentences. The lists were presented in a fixed order using the adaptive SRTn procedure that was selected from experiment 1 (relative keyword scoring with a fixed stepping scheme).

The subjects received written instructions. Their task was to reproduce the sentences by entering them by keyboard in the edit window of the computer program. The sound signal was delivered binaurally by headphones (Philips SBC HP140).

The verification of the responses was done by a simple *one-to-one* string matching algorithm. According to the relative scoring/fixed stepping scheme adaptive procedure as presented in Sec. III, each keyword was scored separately, and the presentation level of the next sentence was adapted dependent on the relative number of keywords that were reproduced correctly. For the remaining part, the test procedure was the same as described in Sec. III.

2. Results

A total of 416 responses were recorded, containing 1876 words. Fifteen of these words contained a typing or spelling error, constituting an error rate of 0.8%. Twelve were typing errors and three were spelling errors. Furthermore, 12 of the 15 errors appeared in a keyword, which means that on average each measurement contained 0.4 false negatives (keyword scored false due to a typing or spelling error). In ten cases the response was a correctly spelled Dutch word, strongly resembling the intended keyword. Examples hereof are "We *konden* nog niet in het gebouw" for "*Honden* mogen niet in het gebouw" and "Er wordt hier geen *lijst* opgebouwd" for "Er wordt in dit land geen *rijst* verbouwd." Following the strict criterion as used by Plomp and Mimpen (1979), these words were counted as an erroneous response and not as a typing or spelling error.

3. Discussion and conclusions

Given the low error rate, it may seem questionable whether it is necessary to have an algorithm that deals with typing or spelling errors. If the error rate was 2.5 times as high, so that each measurement contained one false negative, the effect on the final value of the SRTn could be at most 0.3 dB.[4] However, it is important to rule out the theoretical possibility that the test outcomes can be dependent on the listeners' typing and spelling skills. In other words, the low error rate does not rule out the need for an algorithm.

The errors mainly consisted of typing errors. Furthermore the results confirm that the implementation of the algorithms (choosing and evaluating the parameter values) should focus on the differentiation between typing and spelling errors and true errors that strongly resemble a keyword.

C. Implementation

1. Spelling checker

The spelling checker algorithm that was used is a very straightforward string matching algorithm that consists of three steps. The response is converted into a list of words, and each item on the list is checked in order to determine whether it is a known word in the dictionary. (It was verified that all words marked as keywords in the Plomp and Mimpen sentences were in the dictionary.) If not, the word is replaced by the first variant that is suggested by the spelling checker (e.g., the response *ryst* is not in de dictionary and is replaced by the first suggestion *rist*). Subsequently, each of the reference keywords is matched against the response list by one-to-one string matching. If a keyword occurs in the list, the matching function returns true and the score is incremented. Only one occurrence of a keyword is scored.

2. Dynamic alignment

The basic routine of the dynamic alignment algorithm is based on an algorithm that is used to compare, e.g., protein or DNA sequences. The algorithm first creates a two-dimensional matrix in which the two strings to be matched (the reference keyword and the response word) are set out on the x-axis and y-axis. Subsequently, the matrix is filled by comparing each character on one axis with all of the characters on the other axis. Dependent on the outcomes of the comparison, a score or penalty is obtained: 10 for a match, −4 for a deletion or insertion (gap), −4 for a substitution (mismatch) that is within one distance on the keyboard, and −7 for all other substitutions. In order to reduce the influence of mistakes in the sequence of characters within a word (e.g., *faeture* for *feature*), the first two substitutions with a character that is in the reference keyword are not penalized but considered a match. After the matrix is completely filled, the optimal alignment is determined by tracing back the shortest path across the matrix and summing the scores of the individual steps. A matching index is obtained by dividing the alignment score by the maximal possible score (which is the length of the reference keyword times ten). For example, the best path of the word *task* in reference to the word *token* is 10 (match *t*), −7 (substitute *a* for *o*), −4 (insert *s*), 10 (match *k*), −4 (delete *e*), and −4 (delete *n*), which results in an alignment score of 1 (10−7−4+10−4−4) and subsequently a matching index of 0.02 (1/50). The matching index is consequently compared with a threshold value, the matching criterion. The matching criterion was determined by testing the performance on the words found in Sec. IV B. With an equal error rate[5] of 23%, the optimum matching criterion was found to be 0.767. With a matching index above this value, the responded word is considered to be a match; otherwise, it is a mismatch. In this algorithm, each of the keywords is aligned with each item on the list of responded words in the same manner as described for the spelling checker algorithm.

V. EXPERIMENT 2: EVALUATION OF THE AUTOMATIC PROCEDURE

A. Methodology and materials

1. Stimuli and design

The original set of sentences (Plomp and Mimpen, 1979) was used for the test. The two automatic verification procedures (keyboard responses with either spelling checker or dynamic alignment paradigm) were evaluated against the original procedure of verbal responses and verification by an experimenter. This experiment had two goals: the practical goal of validating the automatic tests against the test procedure with verbal responses and the goal of testing the performance of the two verification algorithms themselves.

2. Subjects

A total of 12 subjects with no reported hearing problems participated in the experiment. Six of them were young people (ages ranging from 21 to 23) who were experienced in using the keyboard, and six of them were elderly people (ages ranging from 63 to 67) who were less experienced with a keyboard. For each subject, the pure-tone thresholds in both ears were tested. The better ear was used for testing. Some subjects were familiar with speech-in-noise testing, but none of the subjects were familiar with the Plomp and Mimpen sentences.

3. Procedure

The ten lists were presented in a fixed order. Each of three conditions was tested with three lists. The conditions were tested in alternate order, according to a 3×3 latin-square design. The first list was presented as a practice list.

The test was administered by a computer program. The sound signal was delivered monaurally by headphones (Sony MDR-7509) through a SBLive! 400 sound module. The noise level was fixed at a comfortable level of 70 dBA. The subjects received written instructions. The subjects' task was to reproduce the sentences by entering their responses into the designated edit window using the keyboard. In the standard test condition, the responses were entered by an experimenter, the subjects' task being to repeat the sentences verbally.

The responses were evaluated with the verification algorithm under test. Each condition used the relative score/fixed stepping scheme adaptive procedure as described in Sec. IV B 1, with one small alteration. In the original procedure, the first sentence of a list was repeated until the listeners' response was entirely correct. To prevent the possibility of an infinite loop caused by a repeated misspelling of one of the keywords, the number of repetitions of each first sentence of a list was limited to 4.

4. Data analysis

The test procedures were evaluated on the estimated SRTn value, the test-retest reliability, and the slope of the discrimination function. We defined the test-retest reliability by the rms of the differences between two SRTn values for each subject, divided by $\sqrt{2}$ [see experiment 1 and Plomp and Mimpen (1979)] for a situation with two measurements per subject per condition. As the present test consisted of three measurements per condition per subject, we took the average of the three possible rms/$\sqrt{2}$ values as the value for test-retest reliability. This procedure enabled us to make a direct comparison with the results of experiment 1.

The spelling checker and dynamic alignment algorithms were also evaluated directly on the *precision*$\times$*recall* ratio. Precision is calculated as the number of hits divided by the sum of the numbers of hits and false alarms. Recall denotes the ratio of the number of hits to the total number of typing or spelling errors. Both ratios are expressed as a percentage.

TABLE III. Mean SRTn (standard deviation in parentheses) for the three test conditions of experiment 2, overall, and differentiated for the young and elderly groups.

	Mean SRTn (SD) (dB)		
Condition	Overall	Young	Elderly
0: verbal repetition, experimenter	−6.83 (1.09)	−7.17 (0.94)	−6.49 (1.15)
1: keyboard, dynamic alignment	−6.99 (1.17)	−7.18 (1.28)	−6.79 (1.03)
2: keyboard, spelling checker	−6.73 (0.98)	−7.01 (0.94)	−6.42 (0.96)

B. Results

1. SRTn

Mean SRTn values in the three test conditions are given in Table III, both overall and differentiated between the groups young and elderly. The overall mean values are based on the average of 36, 35, and 34 individual SRTns for the conditions verbal, keyboard/dynamic alignment, and keyboard/spelling checker, respectively. Each subject was tested with three lists in each condition; however, the data contain three outliers, all of which occurred in the elderly group. The three outlier values are discarded in the analysis and are described in Sec. V C.

Both automatic verification conditions produced virtually the same SRTns as the verbal reference test. A repeated-measure analysis of variance showed no difference among conditions. Mean SRTns of repeated measurements are presented in Table IV. No effect of learning or fatigue was observed.

TABLE IV. Means for the first, second, and third measurements of SRTns in experiment 2. Mean SRTn refers to the value averaged across presentation levels.

Condition	Mean SRTn 1 (dB)	Mean SRTn 2 (dB)	Mean SRTn 3 (dB)
0: verbal repetition/experimenter	−6.72	−7.02	−6.76
1: keyboard/dynamic alignment	−7.01	−6.95	−7.00
2: keyboard/spelling checker	−6.56	−6.90	−6.71
Overall	−6.77	−6.95	−6.82

TABLE V. Mean SRTn, test-retest reliability, and slope of the discrimination function for the three test conditions of experiment 2. Mean SRTn refers to the value averaged across presentation levels; ML SRTn refers to the value obtained with the ML estimation.

Condition	Mean SRTn (dB)	Mean test-retest reliability (dB)	Slope (%/dB)	ML SRTn (dB)
0: verbal repetition/experimenter	−6.83	0.65	14.8	−6.96
1: keyboard/dynamic alignment	−6.99	0.82	13.4	−7.18
2: keyboard/spelling checker	−6.73	0.75	12.5	−6.96

2. Test-retest reliability

In the analysis of the test reliability, the outlier data points of one list were replaced by the average of the values of the remaining two lists. In terms of test-retest reliability, values vary from 0.82 dB for the dynamic alignment condition to 0.65 dB for the reference test (Table V).

In order to check whether the differences were significant, a permutation test was performed. For each condition 20 permutation samples of 10 out of 12 were drawn from the original data, and for each of these samples the test-retest reliability was calculated. An analysis of variance on the samples indicated a significant effect of condition ($p < 0.01$). A subsequent Tukey-HSD *post hoc* test showed that all conditions differ significantly from each other ($p < 0.01$).

3. Discrimination function

An analysis of variance (with subject and condition as fixed factors) on the standard deviation within lists (i.e., the standard deviation of the mean score over the last ten sentences per list) showed no significant effects or interactions, thus indicating that all subjects behaved in the same manner. To see if a learning effect occurred, the first, second, and third measurements were compared for each condition in a series of paired *t*-tests. No learning effect was observed. Therefore, all data were pooled across subjects for the estimation of the discrimination functions.

Discrimination functions are shown in Fig. 2. Slope values and accompanying estimated SRTns are presented in Table V. Slope values range from 12.5%/dB (keyboard, spelling checker) to 14.8%/dB (verbal repetition, experimenter). Besides a slope value, the ML estimation of the discrimination function also yields an estimation of the SRTn. The ML SRTn is very similar to the mean of the SRTns that were estimated by averaging across presentation levels, differences being at most 0.23 dB (see Table V).

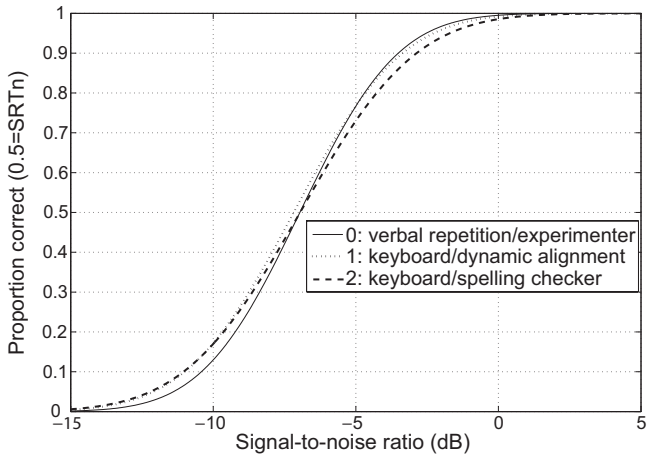

FIG. 2. Discrimination functions for the measurement of the SRTn in the three test conditions in experiment 2. See text and Table V for more details on slope values.

TABLE VI. Algorithm performance as derived from experiment 2.

Algorithm	Precision	Recall	Precision × Recall
1: dynamic alignment	62%	67%	42%
2: spelling checker	100%	29%	29%

4. Algorithm performance

In the keyboard conditions, a total number of 36 keywords contained a typing or spelling error. Fifteen errors occurred in the dynamic alignment condition, of which ten were positively matched. In six cases, a word that was spelled correctly in the context of the response, but was not equal to a keyword, was falsely labeled correct (false alarm). In the spelling checker condition, 21 errors were made, of which 6 were corrected. The spelling checker algorithm produced no false alarms. Precision × recall ratios are 42% and 29%, respectively, for the alignment and spelling checker algorithms (Table VI).

C. Discussion

1. SRTn

The results showed no significant differences in either mean SRTn or mean standard deviation within lists between the young and experienced and the elderly and less experienced keyboard users. Note that a very basic level of typing skill is assumed as people have to know how to type to be able to take the test. However, the results suggest no difference in test results when it comes to age group and self-reported typing skills.

2. Outliers

The three outliers all concern cases in which—due to repeated errors—the first sentence was presented the maximum number of four times before switching to the next sentence. As a result, the second sentence was presented at a SNR of +12 dB. In spite of a series of consecutive correct responses that resulted in presentation levels well below zero further into the procedure, the subjects were unable to reach a normal SRTn.

The outliers indicate that in the current procedure, the repetition of a typing or spelling error in the first sentence is not properly dealt with by the verification algorithm, which can result in a large underestimation of the SRTn. As a solution for this problem, the adaptive procedure was adjusted by allowing the first sentence to be only partially correct (i.e., one keyword is allowed to be incorrect) before switching to the second sentence, instead of requiring the first sentence to be completely correct.

3. Test-retest reliability and discrimination function

The results for the verbal repetition condition are an almost exact replication of the values found in Sec. III. In comparison to the verbal repetition task, both automatic verification conditions have slightly higher test-retest variability. Between keyboard conditions, the results are contradictory. On the test-retest reliability, the spelling checker algorithm performs better than dynamic alignment, whereas in the slope of the discrimination function the dynamic alignment algorithm performs better.

Values reported in the literature are roughly 1–1.5 dB for the test-retest reliability, and 10%–15%/dB for the slope of the discrimination function (group curves; data pooled across subjects). Compared to these values, both automated verification procedures score alike. In light of the reference data that we have found for the original test procedure, the test-retest reliability is maintained (0.65–0.75 dB for the reference, 0.75 or 0.82 for keyboard responses), but the slope has become less steep (21% per dB in experiment 1 to approximately 13% per dB for keyboard conditions in experiment 2). As suggested in Sec. III, the slope of the discrimination function of keyword verification might be improved by renormalizing the sentence material.

4. Algorithm performance

Results show that the dynamic alignment algorithm has the higher precision × recall ratio, although it should be noted that the impact of this observation is limited by the sparsity of errors. Whereas the spelling checker algorithm has a perfect precision, the recall is very low. The error pattern reveals two mechanisms underlying the high number of misses. Either the erroneous response is being replaced by the wrong alternative or the erroneous response is not replaced because it constitutes an existing word. To improve this requires far-reaching adjustments in the dictionary: basically the development of a task specific dictionary or even a set of dictionaries that are keyword specific. This would seriously limit the possibilities of the test being used with other sentence sets.

A closer look at the error pattern of the dynamic alignment algorithm shows an underlying round off problem that causes a lot of false mismatches. As it was implemented in the algorithm, the value of the matching criterion (0.767) is a round off of a fraction (23/30). However, the matching scores are obtained by dividing the score of the shortest path by the maximum score that is possible and are not rounded off. Hence, a word that scored 23/30 (or 46/60) was deemed a mismatch: a score that occurs, e.g., when a six-letter word contains one insertion or deletion. By slightly adjusting the criterion so that these words are scored as a match, the recall of the algorithm is given a boost to 93%. Although this also increases the number of false alarms, this does not affect the precision. As a result, the performance is altered to a precision × recall ratio of 60%.

VI. CONCLUSIONS

In the current paper, we have reported the development of a fully automated procedure for the measurement of the speech reception threshold for sentences in noise, which can be administered by computer. The procedure was based on the test developed by Plomp and Mimpen (1979), which consists of a repetition task of a simple straightforward sentence material that is presented against a background of stationary speech-shaped noise. The results of this study therefore hold for stationary noise and cannot be generalized for test procedures that use fluctuating noise.

The evaluation of different scoring strategies showed that adaptive procedures, which score each keyword separately (relative scoring), have a better test-retest reliability than procedures that score all keywords as a block. Furthermore, adaptive procedures that use a fixed step size show a higher test-retest reliability than procedures that use a converging step size (which decreases as the measurement proceeds). In comparison to the original procedure of whole sentence scoring, the slopes of the discrimination functions are less steep for all procedures that use keywords. However, slope values are still in line with data reported in the literature for other SRTn tests in stationary noise, with small differences between keyword conditions.

Only a limited number of typing and spelling errors were found when collecting responses from listeners using a keyboard. Based on this small data set, dealing with errors by means of a procedure based on a dynamic alignment resulted in less misses and less false alarms than when a simple spelling checker was used. Furthermore, the results indicated that switching from a verbal response task to manually entering the response by keyboard does not affect the test outcomes. In conclusion, the current study shows that it is feasible to obtain accurate SRTn data for sentences using an automated measurement procedure.

ACKNOWLEDGMENTS

This research was supported by grants from the European Union FP6, Project 004171 HEARCOM. The authors wish to thank Gerrit Bloothooft of Utrecht University for his contributions to earlier versions of the manuscript. Niek Versfeld of the Academic Medical Center in Amsterdam and three anonymous reviewers are acknowledged for their critical reading and useful comments.

[1]The entire set of 130 sentences, keywords marked, is available on request.

[2]On first appearance the 3 dB difference between the 1/3 and 2/3 correct scores may seem to be large. The reason behind this is that an adaptive procedure can only be accurate and discriminative if there are enough reversals during the track of the procedure. If a sentence contains only three keywords, the decision about the correctness is based on relatively little information. In such a case a relatively large step size is preferable over a small step size since it is more likely to force a reversal. In the case of 2/5 and 3/5 correct scores, the small difference (2 dB) is justified since it is based on relatively more information.

[3]It should be noted that the curve that was found by Plomp and Mimpen was fitted by eye, whereas we used a ML estimation. This methodological difference might—at least in part—explain the difference between their slope value (15%/dB) and the value that we found (21%/dB) for exactly the same test.

[4]Under the assumption that the up-and-down procedure is the same for the remaining part of the test, one keyword wrong can lead to a difference of at most 3 dB (i.e., an increase instead of a decrease of 1.5 dB) for the next sentence. As the SRTn is calculated by averaging ten presentation levels, this can result in a maximum increase of the SRTn of 0.3 dB.

[5]Equal error rate is generally defined as the error rate where the number of false mismatches and the number of false matches are approximately the same.

Bilger, R., Nuetzel, J., Rabinowitz, W., and Rzeczkowski, C. (**1984**). "Standardization of a test of speech perception in noise," J. Speech Hear. Res. **27**, 32–48.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Gelfland, S., Ross, L., and Miller, S. (**1988**). "Sentence reception in noise from one versus two sources: Effects of aging and hearing loss," J. Acoust. Soc. Am. **83**, 248–256.

Hagerman, B., and Kinnefors, C. (**1995**). "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," Scand. Audiol. **24**, 71–77.

Kalikow, D., Stevens, K., and Elliot, L. (**1977**). "Development of a test of speech intelligibility in noise using sentences with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.

Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am. **102**, 2412–2421.

Letowski, T., Hergenreder, P., and Tang, H. (**1992**). "Relationships between speech recognition threshold, average hearing level, and speech importance noise detection threshold," J. Speech Hear. Res. **35**, 1131–1136.

Macleod, M., and Summerfield, A. Q. (**1990**). "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise," Br. J. Audiol. **24**, 29–43.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Noordhoek, I., Houtgast, T., and Festen, J. (**1999**). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. I. Normal-hearing listeners," J. Acoust. Soc. Am. **105**, 2895–2902.

Noordhoek, I., Houtgast, T., and Festen, J. (**2000**). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. II. Hearing-impaired listeners," J. Acoust. Soc. Am. **107**, 1685–1696.

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, 146–154.

Plomp, R., and Mimpen, A. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

Rooij, J. v., and Plomp, R. (**1991**). "The effect of linguistic entropy on speech perception in noise in young and elderly listeners," J. Acoust. Soc. Am. **90**, 2985–2991.

Smits, C., Kapteyn, T., and Houtgast, T. (**2004**). "Development and validation of an automatic speech-in-noise screening test by telephone," Int. J. Audiol. **43**, 15–28.

Smoorenburg, G. (**1992**). "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearingloss in relation to their tone audiogram," J. Acoust. Soc. Am. **91**, 421–437.

Versfeld, N., Daalder, L., Festen, J., and Houtgast, T. (**2000**). "Method of selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.

Wagener, K., Brand, T., and Kollmeier, B. (**1999a**). "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache I: Design des Oldenburger Satztests [Development and evaluation of a sentence test for the German language I: Design of the Oldenburg sentence test]," Z. Fuer Audiologie, Audiological Acoust. **38**, 4–15.

Wagener, K., Brand, T., and Kollmeier, B. (**1999b**). "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache II: Optimierung des Oldenburger Satzetests [Development and evaluation of a sentence test for the German language II: Optimization of the Oldenburg sentence test]," Z. Fuer Audiologie, Audiological Acoust. **38**, 44–56.

Wagener, K., Brand, T., and Kollmeier, B. (**1999c**). "Entwicklung und Evaluation eines Satztests für die Deutsche Sprache III: Evaluation des Oldenburger Satztests [Development and evaluation of a sentence test for the German language III: Evaluation of the Oldenburg sentence test]," Z. Fuer Audiologie, Audiological Acoust. **38**, 86–95.

Wagener, K., Josvassen, J., and Ardenkjaer, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.

Wagner, R., and Fisher, M. (**1974**). "The string to string correction problem," J. Assoc. Comput. Mach. **21**, 168–173.

Yang, Q., Yuan, S., Zhao, L., Chun, L., and Peng, S. (**2003**). "Faster algorithm of string comparison," Pattern Anal. Appl. **6**, 122–133.